Iris ESHKOL
Université d'Orléans, LLL
E-mail : iris.eshkol@univ-orleans.fr

A quelles conditions une taxinomie du lexique est-elle possible en TAL ?

1. Introduction

Le Traitement Automatique du Langage (TAL) est un domaine pluridisciplinaire qui met en jeu la linguistique et l'informatique. La confrontation de la machine et du langage pose des problèmes théoriques et pratiques considérables, du fait de la complexité de la compétence langagière. Le langage possède certaines propriétés qui le rendent par nature résistant au traitement informatique :

- l'ambiguïté : contrairement aux langages artificiels, le langage naturel est constitué de formes susceptibles d'interprétations différentes ;
- la productivité : la capacité de langage nous permet de produire et de comprendre une infinité de mots et de phrases, de dire une même chose de façons infiniment variées ;
- l'implicite : les échanges langagiers reposent sur des informations implicites, qui font partie des connaissances partagées par les interlocuteurs ou du contexte de communication.

Ces difficultés ont rendu pour l'instant illusoire l'ambition de faire simuler nos capacités langagières par une machine, étant donnée la masse d'informations à prendre en compte et la complexité des mécanismes qui sont à l'œuvre.

Les années 1990 ont vu l'évolution du TAL avec, d'une part, la constitution et l'exploitation de corpus qui ont provoqué une redéfinition des objectifs et un renouvellement des méthodes de la linguistique (Habert et Nazarenko, 1997). L'apparition d'Internet a offert un accès à des données textuelles massives et les outils élaborés par le TAL, par exemple les lemmatiseurs ou les analyseurs syntaxiques, en ont permis une exploitation immédiate. Le deuxième phénomène est la lexicalisation des traitements qui s'explique peut être par le rôle central que le lexique joue dans la gestion des ambiguïtés et des exceptions, deux problèmes majeurs du TAL (Nazarenko, 2006). Beaucoup d'ambiguïtés peuvent être résolues au niveau lexical par des contraintes d'incompatibilités ou de préférences lexicales et nombre d'exceptions peuvent être décrites dans le lexique. La question des ressources lexicales et de leur disponibilité prend donc une grande importance surtout lorsqu'il s'agit du traitement des corpus.

On distingue différents types de ressources :
- lexiques, listes d'unités linguistiques simples ou composées ;
- dictionnaires, unités linguistiques avec des définitions ;
- thésaurus, listes d'unités qui explicitent en plus les relations que ces dernières entretiennent entre elles ;
- ontologies, structures hiérarchiques de concepts.

Dans cet article, nous montrerons l'exploitation du modèle taxinomique dans la représentation du lexique pour le TAL et nous essaierons, en conclusion, de répondre à la question du titre : à quelles conditions une taxinomie du lexique est-elle possible en TAL ? Dans la première

partie, nous nous arrêterons sur les lexiques utilisés au cours de l'étiquetage morpho-syntaxique du corpus oral. Nous montrerons comment les décisions d'affectation des items à une catégorie grammaticale requièrent d'être reconsidérées en fonction des critères hétérogènes de classement. Dans un deuxième temps, nous étendrons la question aux dictionnaires et plus précisément aux dictionnaires de prédicats. Nous exposerons différents critères linguistiques qui permettent de délimiter une classe de prédicats telles que les propriétés de sous-catégorisation, de variantes paraphrastiques, de sémantique lexicale (hyperonymie/hyponymie, synonymie/antonymie, etc.) et les propriétés aspectuelles. La démonstration s'appuiera sur l'exemple des prédicats de parole.

2. Lexiques ou étiquetage morpho-syntaxique de l'oral

2.1. Définition de l'étiquetage

L'étiquetage morpho-syntaxique d'un texte est une étape fondamentale de son analyse, et un préliminaire à tout traitement de plus haut niveau. L'objectif est d'attribuer à chacun des mots d'un corpus une étiquette qui récapitule ses informations morpho-syntaxiques. Ce processus peut s'accompagner de celui de la lemmatisation, dont l'objectif est de ramener l'occurrence d'un mot donné à sa forme de base ou « lemme » :

Mot du corpus	Lemme	Etiquette morpho-syntaxique
comment	comment	ADV (adverbe)
vous	vous	PPER2P (pronom personnel 2ème personne pluriel)
faites	faire	VINDP2P (verbe indicatif présent 2ème personne pluriel)
vous	vous	PPER2P (idem)
une	un	DETIFS (déterminant indéfini féminin singulier)
omelette	omelette	NCFS (nom commun féminin singulier)

Deux approches sont en concurrence. L'une, fondée sur les dictionnaires, consiste à décrire aussi exhaustivement que possible l'ensemble des mots du lexique ; l'autre repose sur des règles de calcul morphologique. Néanmoins, un certain nombre de problèmes doit être résolu au cours du processus d'étiquetage.

2.2. Problèmes de l'étiquetage

L'étiquetage morpho-syntaxique considère les formes du corpus une à une, sans prendre en considération les contextes d'apparition. Sa principale difficulté est due à l'ambiguïté des mots polycatégoriels :

> *et vous êtes pour ou contre* (*contre contrer VINDP3S* à la place de *contre contre PREP*[1])

[1] Les exemples sont issus de l'étiquetage d'ESLO par l'étiqueteur Cordial.

La prise en compte du contexte permet de résoudre ce type de problèmes. Nous mentionnerons, au nombre des méthodes de désambiguïsation :

- les dictionnaires des formes de langue ;
- les méthodes à base de règles ;
- les systèmes probabilistes reposant sur des chaînes de Markov ;
- les arbres de décision.

Le contexte concerne l'entourage lexical proche (co-occurrence) droite/gauche de l'unité lexicale, à savoir, les deux ou trois mots qui précèdent et qui suivent.

Les étiqueteurs doivent aussi faire face à des mots non reconnus par leurs dictionnaires : mots erronés ou mal orthographiés, noms propres, néologismes, ainsi qu'à des locutions :

en en PREP (préposition)
effet effet NCMS (nom commun masc. sing.)

Une locution adverbiale *en effet*, dans cet exemple, est décomposée en deux unités étiquetées chacune séparément au lieu d'avoir une étiquette pour tout l'ensemble :

en effet ADV (adverbe)

Cependant, le problème majeur provient, selon nous, de la différence entre les étiqueteurs eux-mêmes. En effet, la diversité en taille et en visée des jeux d'étiquettes attribuées et des stratégies d'étiquetage sous-jacentes diffèrent d'un étiqueteur à l'autre. Les jeux d'étiquettes ne sont généralement pas comparables car les différences persistent à de nombreux niveaux, notamment le nombre d'étiquettes, leurs appellations et limites de profondeur. Et même si des étiquettes sont identiques, l'extension de ces étiquettes peut être très différente. Le problème est particulièrement aigu pour les catégories fermées, déterminants, pronoms, adjectifs indéfinis, etc. où l'on rencontre des différences d'appréciation quant au placement des mots dans les catégories. Ainsi, à l'intérieur d'un même système où l'on distinguerait des "déterminants" et des "numéraux", on peut prendre plusieurs décisions concernant le mot *un* : on peut le considérer comme un déterminant et un numéral, ou bien comme un déterminant et pas un numéral :

qui est <u>un</u> des plus beau château,

A cela s'ajoutent des divergences théoriques. On peut ainsi, par exemple, avoir des étiquettes pour les articles, et considérer les possessifs (*mon, ton, son,* etc.) comme faisant partie des adjectifs, ou bien inclure les uns et les autres dans une catégorie "déterminants". Il faudrait reconnaître également l'absence de théorie bien claire pour un certain nombre de phénomènes (voir tout le domaine des adjectifs indéfinis en français, par exemple). Dans de nombreux étiqueteurs, la théorie linguistique sous-jacente est très rustique, dans d'autres, de nombreux cas d'étiquetage considérés comme satisfaisants dans beaucoup d'applications d'ingénierie semblent assez critiquables sur le plan linguistique. Ces différences proviennent de causes multiples : différences des principes des étiqueteurs, différences d'applications pour les textes étiquetés, etc. Ainsi, selon Jean Véronis et Liliane Khouri (1995),

les étiqueteurs probabilistes sont très sensibles au jeu d'étiquettes qu'on leur donne. Trop grossier, le jeu d'étiquettes ne permet pas de capturer assez de propriétés distributionnelles à travers les transitions markoviennes. Trop fin, il impose des tailles de corpus gigantesques pour avoir un échantillon suffisant de transitions observées lors de l'apprentissage (en particulier lors de l'utilisation de trigrammes)[2] .

Une difficulté supplémentaire apparaît dans le cadre multilingue, due au fait que les phénomènes morpho-syntaxiques que l'on cherche à représenter par des étiquettes ne sont pas forcément les mêmes dans les différentes langues.

Pour démontrer nos propos, nous présenterons des jeux d'étiquettes conçus par deux outils différents. Le premier, TreeTagger[3], a été développé par Helmut Schmid à l'Université de Stuttgart. Son étiquetage est basé sur les méthodes probabilistes (Schmid, 1994).

ABR abréviation
ADJ adjective
ADV adverbe
DET:ART, DET:POS (ma, ta, etc.)
INT interjection
KON conjonction
NAM nom propre, NOM nom
NUM numérique
PRO pronom, PRO:DEM, PRO:IND, PRO:PER, PRO:POS (mien, tien, etc.),
PRO:REL
PRP préposition, PRP:det (au, du, aux, des)
PUN ponctuation, PUN:cit (citation), SENT phrase
SYM symbole
VER:cond, VER:futu, VER:impe, VER:impf, VER:infi, VER:pper, VER:ppre,
VER:pres, VER:simp, VER:subi, VER:subp

Figure 1 : jeu d'étiquettes de TreeTagger pour le français

Le classement établi par TreeTagger (Figure 1) contient neuf catégories « classiques » grammaticales comme *nom, adverbe, adjectif*, et quatre supplémentaires : *abréviation, numérique, ponctuation* et *symbole*. Les catégories *NUM, SYM* et *PUN* peuvent être reconnues grâce aux dictionnaires de l'outil où ces unités sont énumérées. La reconnaissance des *ABR, SYM et PUN* peut faciliter la tâche de la segmentation du texte en phrases et en unités distinctes, tâche qui précède celle de l'étiquetage après qu'on a appliqué les règles de leur désambiguïsation. En effet, les signes de ponctuation comme *point* ou *tiret* peuvent faire partie des mots composés (*chou-fleur)* ou des abréviations (*S.N.C.F.)* sans parler des noms de civilité (*M.Dupont)* ou des appellations de fichiers (*rapport.doc).* En ce qui concerne la profondeur d'étiquettes, ces dernières ne prennent pas en compte les indications morphologiques sur le genre, le nombre, la personne. Il semble que les informations proposées qui ne marquent que les parties du discours (nom, adjectif, etc.), leur type (pronom

[2] Véronis, J., Khouri, L. (1995). *Etiquetage grammatical multilingue: modèle.* Document MULTEXT LEX2 (http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX2_1.html)
[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

démonstratif, personnel, possessif, etc.) ou leur temps et mode (verbe conditionnel, imparfait, etc.) garantissent moins d'erreurs.

Le deuxième outil est Cordial. Ce logiciel est développé par l'entreprise Synapse. C'est un étiqueteur du français écrit, avec une large palette d'étiquettes :

Adjectif
(Genre + Nombre) / Invariant en (Nombre et/ou Genre) + Débutant par un "h" aspiré
Démonstratif / Possessif / (Numérique + Cardinal / Ordinal)
Interrogatif / Indéfini
Adverbe
Article
(Défini / Indéfini) + ((Genre + Singulier) / Pluriel Invariant en Genre)
Conjonction
de Coordination / de Subordination
Interjection
Préposition
Nom
(Genre + Nombre) / Invariant en (Genre et/ou Nombre) + Débutant par un "h" asp
Pronom
Personnel : Personne + Nombre
Démonstratif : Nombre
Relatif : (Genre + Nombre) / Invariant (en Genre et en Nombre)
Indéfini : (Genre + Nombre)/Invariant en (Genre et/ou Nombre)
Possessif
Verbe
Mode + Temps + Personne + Nombre
Ponctuation Faible / Forte (fin de phrase)

Figure 2 : jeu d'étiquette de Cordial (présentation récapitulative)

Cordial utilise environ 200 étiquettes indiquant les différentes informations morphologiques comme le genre, le nombre ou l'invariabilité pour les noms et les adjectifs ; la distinction en mode, en temps et en personne pour les verbes ; et même la présence du h aspiré au début du mot.

En comparant seulement deux outils, nous avons déjà pu constater un grand nombre de différences aux niveaux qualitatif et quantitatif liées, selon nous, aux choix applicatifs et aux techniques de développement des concepteurs des outils.

> Il n'y pas de meilleur jeu d'étiquettes, […] dans la pratique la plupart des jeux d'étiquettes constituent plutôt des compromis entre la finesse de la description linguistique et ce qui peut être attendu, pour des raisons pratiques, d'un système automatique d'étiquetage. (Leech, 1994 : 15).[4]

2.3. Classements non adaptés à l'oral

L'oral rend la tâche d'étiquetage encore plus difficile. Des étiqueteurs conçus pour les textes écrits sont mal adaptés aux spécificités d'une langue moins normalisée. En effet, les

[4] Leech, G. (1994). « 100 million words of English : the British National Corpus ». *English Today* 9(1) : 9-15.

transcriptions de corpus oraux ne sont en général pas ponctuées pour éviter l'anticipation de l'interprétation (Blanche-Benveniste et Jeanjean, 1987). Les signes de ponctuation comme le point ou la virgule, ainsi que la majuscule au début de l'énoncé, sont des marques typographiques. De même, la notion de phrase, essentiellement graphique, a rapidement été abandonnée par les linguistes qui s'intéressent à l'oral. Les études sur la langue parlée ont permis ensuite de dégager des phénomènes propres à l'oral, qu'on regroupe souvent sous l'appellation générale de *disfluences* : répétitions, autocorrections, amorces de mots, etc. Le corpus ESLO[5], auquel nous nous intéressons dans cet article, provient de la transcription d'enregistrements oraux, et présente donc des particularités mal prises en compte par les étiqueteurs standard. En accord avec Blanche-Benveniste (2005), nous considérons que les disfluences doivent être intégrées par l'analyse linguistique même si elles créent des difficultés pour le traitement.

Citerons quelques erreurs propres à la nature orale des données que nous avons rencontrées au cours de l'étiquetage d'ESLO par l'étiqueteur de l'écrit Cordial :

- troncation ou amorce : dans les conventions d'ESLO, la séquence amorcée est notée par un tiret ce qui pose évidemment un problème dans l'étiquetage :

on fait une ou deux réclam- réclamations
réclam- réclamations réclamréclamations NCMIN[6]

au lieu d'être analysée en deux unités séparées :

réclam- reclam- NCI[7]
réclamation réclamation NCFS

- interjection :

alors ben (ben ben NCMIN) *écoutez madame*

Ce phénomène pose le problème de l'ambiguïté car, selon Dister (2007)

> Toute forme peut potentiellement devenir une interjection. On assiste alors à une recatégorisation grammaticale […], le phénomène par lequel un mot ayant une classe grammaticale dans le lexique peut, en discours, changer de classe (p. 350).[y1][8]

[5] L'Enquête SocioLinguistique d'Orléans (ESLO) a été menée entre 1968 et 1971 par des professeurs de français de l'University of Essex (Royaume-Uni) et avait pour but de récolter des documents sonores dans une visée didactique. ESLO représente un corpus oral de grande taille : 317 heures de paroles spontanées et comporte des fiches sur plus de 200 locuteurs. Les situations d'enregistrements sont diverses : des entretiens en face à face, des discussions entre amis, des enregistrements en micro caché, des interviews de personnalités de la ville, des conférences ou débats ainsi que des entretiens au Centre Médico Psychopédagogique d'Orléans (entretiens entre une assistante sociale et des parents). En 2005, le laboratoire CORAL devenu ensuite LLL (Laboratoire Ligérien de Linguistique) a entrepris de mettre à disposition ce corpus dans le respect des méthodes et des techniques actuelles et de constituer une nouvelle enquête ESLO 2 comparable à ESLO 1. Réunis, ESLO 1 et ESLO 2 formeront une collection de 700 heures d'enregistrement.

[6] Nom commun masculin invariant en nombre

[7] Nom commun invariable

[8] Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque*

j'ai quand même des attaches euh ben de la campagne qui est proche quoi (PRI[9])

Il en va de même d'autres éléments, comme *hein, bon, bien, quoi, voilà, comment dire, etc.* qui apparaissent avec une fréquence élevée dans les corpus oraux et qui, sans ponctuation, peuvent être ambigus. Ces mots constituent des énoncés à eux seuls ou se manifestent à différentes places d'un énoncé sans intégrer sa structure (c'est-à-dire sans entrer en relation syntaxique avec un autre élément).

- répétition et autocorrection :

  je crois que le (le le PPER3S au lieu de le le DETDMS) *le* (le le DETDMS) *les saisons*

Il faut noter également un certain nombre d'erreurs provenant des fautes de frappe ou d'orthographe faites par les transcripteurs humains, les transcriptions n'ayant pas été soumises aux correcteurs orthographiques. La correction manuelle d'un fichier étiqueté par Cordial Analyseur a permis d'établir approximativement le taux d'erreur réalisé par le logiciel à 4%.

2.4. Jeu d'étiquettes proposé

Afin de mieux adapter l'étiquetage à nos besoins, un certain nombre de modifications ont été apportées au jeu d'étiquettes de Cordial. D'une part, nous avons essayé de réduire le nombre d'étiquettes tout en gardant les informations nécessaires, selon nous, à l'analyse linguistique. D'autre part, nous avons été obligée d'adapter les étiquettes à notre corpus et aux conventions de sa transcription :

Adjectif
(Genre + Nombre) / Invariant / Numérique (ordinaux)
Adverbe
Déterminant
(Défini / Indéfini / Démonstratif / Possessif / Interrogatif) + ((Genre + Nombre) / Invariant)
Conjonction
de Coordination / de Subordination
Interjection
Préposition
Nom
(Commun/Propre) + (Genre + Nombre) / Invariant
Pronom
Personnel : Personne + Nombre+Genre
Démonstratif : Genre+Nombre
Relatif : (Genre + Nombre) / Invariant
Indéfini : (Genre + Nombre)/Invariant
Possessif : Personne+Genre + Nombre
Interrogatif : (Genre + Nombre)/Invariant

---

de données textuelle orale VALIBEL, Thèse de Doctorat, Université de Louvain.
[9] Pronom relatif invariable

Verbe
Mode + Temps + Personne + Nombre
Mot inconnu
Présentateur
Chiffre

Figure 3 : jeu d'étiquettes proposé (Récapitulatif)

De nouvelles étiquettes ont été introduites comme MI (mot inconnu) pour, entre autres, les cas de troncations et PRES (présentateur) pour les tournures comme *il y a, c'est, voilà* très présentes à l'oral. Quelques étiquettes, trop détaillées selon nous dans Cordial, ont été simplifiées. Par exemple, la gamme d'étiquettes concernant les invariances de l'adjectif ou du nom (masculin invariant en nombre, féminin invariant en nombre, singulier invariant en genre, pluriel invariant en genre, invariant en nombre et en genre) a été réduite à une seule étiquette (invariable). Par ailleurs, les étiquettes concernant le trait du h aspiré au début du mot ont été supprimées. Afin d'uniformiser le système, certaines étiquettes ont été enrichies : par exemple, les indications sur le genre et le nombre ont été ajoutées aux déterminants démonstratifs et possessifs par souci de cohérence avec d'autres types de déterminants définis ou indéfinis. Cordial ne reconnaît pas toutes les interjections présentes dans le corpus oral, nous avons donc élargi la liste pour cette catégorie.

Nous avons construit par la suite un étiqueteur par apprentissage automatique, à partir de données étiquetées par Cordial et corrigées à la main. Pour faciliter le traitement informatique et améliorer le résultat de l'étiquetage par apprentissage automatique, nous avons proposé de structurer les étiquettes sur 3 niveaux, appelés respectivement L0 (niveau des étiquettes sur les parties du discours), L1 (niveau des variantes morphologiques) et L2 (niveau syntaxico-sémantique) (Figure 4) :

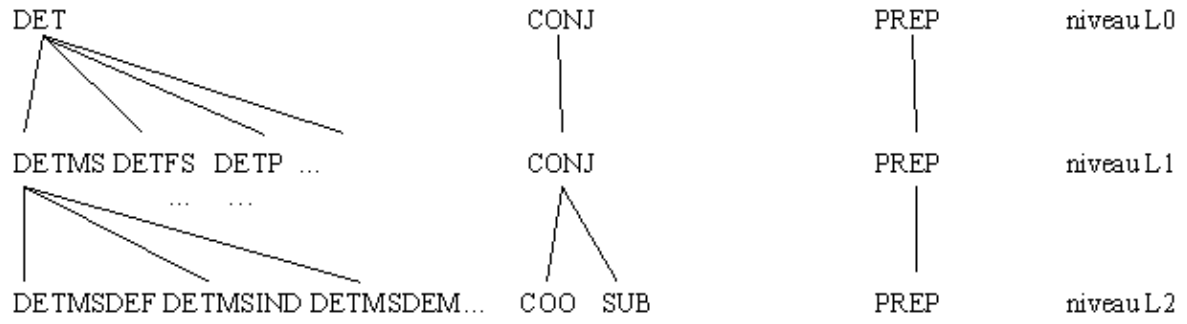

Figure 4 : structuration hiérarchique de quelques étiquettes

Certaines étiquettes restent les mêmes sur les 3 niveaux, d'autres ne changent qu'au deuxième, comme les pronoms et déterminants, et les dernières enfin intègrent chaque fois de nouvelles informations.

Pour améliorer les performances d'un système d'apprentissage automatique, nous avons exploité d'autres connaissances linguistiques provenant de la morphologie flexionnelle dont nous ne parlerons pas ici.

3. Dictionnaire de prédicats

Dans cette partie, nous proposons un exemple de dictionnaire, un autre type de ressources lexicales. Nous essaierons de tenir compte des informations syntaxiques, collocationnelles, paradigmatiques et sémantiques de prédicats[10] sans pour autant les décrire individuellement, comme le font les dictionnaires classiques, mais en les regroupant en classes. En nous fondant sur la théorie des classes d'objets de Gross (1994), le modèle sens-texte de Mel'čuk (1992) et la proposition de norme des lexiques pour le traitement automatique du langage du groupe de travail de l'AFNOR « Lexique pour le TAL »[11], nous proposons un modèle de représentation d'une classe de prédicats de parole.

3.1. Classe de prédicats

Le point de départ de cette démarche est la relation fondamentale entre le prédicat et ses arguments. Le prédicat se définit par sa construction syntaxique et par le choix lexical des arguments qu'il admet. La délimitation du domaine d'arguments d'un prédicat est donc une tâche primordiale.

L'idée de départ est que, d'une part, les prédicats ayant le même sens ou un sens proche et/ou provenant de la même racine ont une distribution similaire. Observons, par exemple, les prédicats comme *ordonner, commander, conseiller,* qui tous désignent l'acte de parole où le locuteur exprime sa volonté, son désir que son interlocuteur fasse quelque chose, d'où la possibilité pour tous ces prédicats d'être suivis d'un infinitif et d'une complétive au subjonctif. Pourquoi faudrait-il séparer le prédicat verbal *demander* du prédicat nominal *demande* actualisé à l'aide d'un verbe support *faire* (*faire une demande*), si ces deux prédicats désignent le même acte de parole ? Ainsi, on crée les classes de prédicats comportant les différents types de prédicats sémantiquement homogènes, dont les arguments sont en nombre identique et ont la même nature.

Cette vision pourrait résoudre le problème de l'ambiguïté. En effet, il semble que deux sens différents d'un prédicat n'ont jamais les mêmes schémas d'arguments. Par exemple, le verbe « dire » n'aura pas le même sens s'il est suivi d'une complétive ou d'un infinitif :

dire à qqn que - usage assertif
dire à qqn de faire qqch - usage directif

Si l'on veut discriminer les différents sens d'un prédicat donné, on doit être en mesure de définir chacune de ses positions argumentales. On construit, par conséquent, des classes sémantiques qui définissent les différents emplois d'un prédicat donné.

La représentation lexicale en groupe où les mots sont classés selon les principes d'homogénéité sémantique et syntaxique, permet de faciliter le traitement automatique. La description se fait, dès lors, à partir d'un ensemble et non plus d'éléments. Dans ces conditions, on procède à une mise en facteur commun des propriétés de tous les éléments d'une classe donnée.

[10] Le prédicat est considéré ici comme un mot qui sélectionne ses arguments.
[11] http://www.normalangue.org

3.2. Prédicats de parole

Les prédicats sur lesquels nous avons travaillé sont les prédicats de parole. On peut en distinguer, en premier lieu, quatre grandes classes :

1. prédicats qui n'indiquent que l'acte de parole lui-même et la personne qui parle :
   parler, bégayer, etc.
2. prédicats qui impliquent en plus l'interlocuteur :
   interpeller, insulter qqn
3. prédicats qui mentionnent l'interlocuteur et le contenu d'un acte de parole, mais en l'évoquant, sans préciser les mots dits ; ces prédicats n'indiquent que le thème, le sujet dont on parle :
   parler de qqch à qqn
4. prédicats qui mentionnent le contenu des propos, ce qui est dit, et seront classés selon ce contenu :
   dire à qqn une information inconnue,
   dire un texte
   dire qqch de caché
   dire ses motifs
   dire ses sentiments

La typologie des prédicats de parole établie a mis en évidence l'importance de certains phénomènes pour la description de l'unité lexicale : aspect, intensité, itérativité, réciprocité, etc. L'acte de parole peut avoir lieu une ou plusieurs fois et peut dépasser les mesures d'un acte de parole ordinaire. L'intensité de l'acte de parole se manifeste dans :

- la quantité de propos prononcés :

une collection, une foule, une bordée, une pelletée d'injures

- la durée

parler infiniment

- la façon de dire

demander avec insistance, avec instance, avec force
crier / murmurer

Les prédicats de parole peuvent renvoyer vers un acte itératif :
répéter, redire qqch à qqn.

De la même façon, l'acte de parole peut commencer, durer, se terminer, ainsi il varie également selon les nuances aspectuelles.

Ces phénomènes peuvent être intrinsèques, propres à chaque prédicat. Dans ce cas, on ne peut pas lui ajouter d'autres marqueurs du même aspect pour des raisons de redondance :

**répéter des radotages*

La valeur aspectuelle peut aussi être ajoutée au prédicat neutre par d'autres termes marqués aspectuellement :

répéter un discours

Nous appelons ces accompagnateurs des termes appropriés aux prédicats. Cette notion est proche de celle des fonctions lexicales de Mel'čuk (1992). La différence est que les termes appropriés d'un prédicat ne sont pas toujours imprévisibles. Tous ces facteurs doivent faire partie de l'article du dictionnaire décrivant le prédicat.

3.3. Modèle proposé

Pour regrouper les prédicats dans la même classe, nous avons suivi les principes suivants :

- la construction de chaque classe de prédicats est faite à partir d'une identité sémantique commune aux trois types de prédicats : prédicat verbal, nominal et adjectival.
- tous les éléments de la classe doivent présenter les mêmes caractéristiques syntaxiques et sémantiques.
- le nom de la classe est un mot vedette faisant partie de la classe, un terme générique, ou bien c'est le mot qui exprime le mieux la valeur sémantique de la classe.
- ce mot vedette est suivi des divers synonymes classés en fonction de nuances de plus en plus particulières.
- les termes appropriés caractérisent les éléments d'une classe et ajoutent des nuances stylistiques et aspectuelles.

Le modèle que nous proposons est composé de deux parties : la description de la classe, c'est-à-dire des propriétés communes à tous les prédicats de la classe et la description de chaque prédicat qui fait partie de cette classe.

La classe est caractérisée par son identifieur (un chiffre qui précise sa position par rapport aux autres classes : classes supérieures ou classes sœurs) et son nom. On indique ensuite les informations concernant la définition sémantique de la classe et la structure argumentale des prédicats. Dans la définition sémantique, on précise l'idée sémantique qui réunit tous les prédicats de la classe. Dans la structure argumentale, on indique le nombre maximal d'arguments que peuvent sélectionner les prédicats de la classe et on décrit chaque argument en mentionnant sa position syntaxique (premier, deuxième argument), sa nature/son type (GN, V-inf), son introducteur (s'il est introduit par une préposition), le mode (indicatif, subjonctif, etc. s'il s'agit de la proposition complétive), son rôle sémantique dans la structure (agent, etc.) et la classe d'objets auquel il appartient.

La deuxième étape est la description de chaque prédicat de la classe. L'entrée lexicale de chaque prédicat comporte trois zones de description : sémantique, elle donne une définition du sens de l'entrée ; syntaxique, elle contient des informations sur le comportement syntaxique du mot ; enfin la zone de termes appropriés au prédicat. Un prédicat est constitué par son domaine d'arguments. Nous noterons, dans chaque cas, la suite des arguments la plus longue. On précise le type du prédicat (prédicat nominal, verbal, etc.) et on note les

informations propres à chaque type : un verbe support pour un prédicat nominal, des transformations et des restructurations pour un prédicat verbal, etc. Ainsi, les prédicats verbaux de la classe <ORDONNER> permettent la transformation d'une complétive en un infinitif sous condition d'une coréférence entre le deuxième argument (l'interlocuteur - N1h[12]) du prédicat et le sujet N0 de la complétive :

N0 ordonne/interdit/permet à N1h$_{(i)}$ que N0$_{(i)}$ fasse qqch $\Rightarrow$
N0 ordonne/interdit/permet à N1h de faire qqch

Pour achever la description, on énumère les termes appropriés de chaque prédicat (voir la Figure 5).

```
<?xml version="1.0" ?>
- <classEntrySystem id="10-1-1-5" name="ORDONNER">
  - <senseClass>
      <definition>un acte de parole par lequel un chef, une autorité manifeste sa volonté ; ensemble de dispositions impé
        caractérisée par une certaine intonation.</definition>
      <comment>«can be represented in terms of the sentence «I want you to do it», and the additional semantic compo
        and «I assume that you have to do what I say I want you to do» (A.Wierzbicka 1987 : 11-12) «Quand on ordonne
        [...]» (D.Vanderveken 1988 : 186).</comment>
    </senseClass>
  - <argumentsClass number="3">
    - <argumentStructure>
        N0 PRED à|de N1 de V-INF
      - <argument num="0" type="GN">
          <definition>un chef, une autorité, prétend l'être</definition>
          <objectsClass>C:HUMAIN_AUTORITE</objectsClass>
        </argument>
      - <argument num="1" type="GN" introducer="à|de">
          <definition>celui à qui on ordonne, est subordonné, est un interlocuteur</definition>
          <objectsClass>C:HUMAIN</objectsClass>
        </argument>
        <argument num="2" type="V-inf" introducer="de" />
      </argumentStructure>

    </argumentsClass>
  - <lexicalEntry id="ID">
      <predicat type="verbal">ordonner</predicat>
    - <sense>
        <definition>"Prescrire par un ordre" (Petit Robert)</definition>
        <example>Je vous ordonne de vous taire. Il ordonne que tout le monde soit convoqué chez lui.</example>
      </sense>
    - <syntacticSystem>
        N0 ordonne à N1 de inf N0 ordonne que P + subj
        <restructurations>N0 ordonne à N1 de V-inf = N0 ordonne que N1 + V-subj</restructurations>
      </syntacticSystem>
    - <appropriateTerms>
      - <term type="Adv">
          expressément
          <example>ordonner expressément à (qqn)</example>
          <aspetualStylisticalNote intensity="high" />
        </term>
      </appropriateTerms>
    </lexicalEntry>
  </classEntrySystem>
```

Figure 5

3.4. Quelques dysfonctionnements

Le modèle proposé a beaucoup d'avantages. Cependant, l'homogénéïté demandée n'est pas toujours possible à réaliser. Dans la classe <jurons et blasphèmes>, les deux prédicats *jurer* et *blasphémer*, n'ont pas une structure d'arguments identique, malgré leur similitude sémantique apparente :

N0 jure ?contre N1n/ N0 blasphème contre N1n

Le même problème se pose pour la classe <déballer>, sous-classe de « dire qqch de caché » :

[12] N0 est un premier argument nominal du prédicat, N1 est le deuxième, etc. Nh est un argument de nature humaine.

N0 confie / livre N2n = secret, sentiments à N1h
N0 confie ce que P, que P à N1h

Il reste aussi quelques prédicats singletons :

saluer qqn
présenter qqn à qqn (?faire la présentation de qqn à qqn)

Enfin, on ne trouve pas toujours d'équivalent entre le prédicat verbal et le prédicat nominal :

faire des salamalecs, politesses, civilités à N1h

4. Conclusion

Pour être modélisées, des connaissances linguistiques devraient être présentées d'une manière exhaustive, d'une part, mais aussi homogène, d'autre part, l'homogénéité étant la condition indispensable si l'on veut éviter les problèmes posés par l'ambiguïté. Quel est le format le plus adapté pour représenter ces connaissances ? Dans cet article, nous avons proposé deux exemples d'utilisation de taxinomie pour représenter les ressources lexicales : le classement du lexique pour l'étiquetage morpho-syntaxique et le dictionnaire de prédicats de parole. Dans les deux cas, il s'agit d'un classement censé être homogène. Dans les deux cas, nous avons essayé de privilégier une structure arborescente qui rend, semble-t-il, le traitement automatique plus efficace. Pour que la taxinomie réussisse, il faut que le domaine de son application soit limité et bien précis. Plus on limite la portée de taxinomie, plus elle semble devenir homogène et exhaustive et donc plus facilement traitable par la machine.

Blanche-Benveniste, C. (2005). « Les aspects dynamiques de la composition sémantique de l'oral ». *Sémantique et corpus.* Lavoisier, Hermes, Paris, 39-74.

Blanche-Benveniste, C., Jeanjean, C. (1987). *Le Français parlé. Transcription et édition*, Didier Érudition, Paris.

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL,* Thèse de Doctorat, Université de Louvain.

Eshkol, I. (2002). *Typologie sémantique des prédicats de parole.* Thèse de doctorat.

Eshkol I., Tellier I., Taalab S., Billot S. (2010), « Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques », *10th International Conference on statistical analysis of textual data (JADT 2010),* Rome, Italie.

Eshkol, I. (à paraître). « Interpréter le contexte dans un corpus oral : fonctions et limites du traitement automatique des données linguistiques ». *Rencontres Interdisciplinaires sur les Systèmes Complexes Naturels et Artificiels,* Rochebrune, France.

Gross, G. (1994). « Classes d'objets et description des verbes ». *Langages*, 115 : 15-30.

Habert, B., Nazarenko, A. (1997). *Les linguistiques de corpus*, A. Colin, Paris.

Leech, G. (1994). « 100 million words of English : the British National Corpus ». *English Today* 9(1) : 9-15.

Mel'čuk, I. (1992). *DEC, Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III.* Les Presses de l'Université de Montréal, Canada.

Nazarenko, A. (2006). « Le point sur l'état actuel des connaissances en TAL ». *Compréhension des langues et interaction*, Lavoisier, Hermes, Paris, 31-70.

Schmid, H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees.* Proceedings of International Conference on New Methods in Language Processing, Manchester. (http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf)

Tellier I., Eshkol I., Taalab S., Prost J-P., (2010). "POS-tagging for Oral Texts with CRF and Category Decomposition", *Research in Computer Science, special issue : Natural Language Processing and its Applications*, 79-90.

Véronis, J., Khouri, L. (1995). *Etiquetage grammatical multilingue: modèle*. Document MULTEXT LEX2 (http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX2_1.html)